Regression in Python

Test name: Regression in Python

Dataset: D:\DATA ANALYSIS\H Regression Tests and Models\Regression in Python\dataset.csv

Original rows: 649

Rows used: 649

Outcome: G3

Simple model: G3 ~ G2

Multiple model: G3 ~ G1 + G2 + studytime + failures + absences + age + Medu + Fedu

Simple regression results:

R-squared = 0.843730

Adjusted R-squared = 0.843489

F = 3493.281566

Model p-value = 5.64240149e-263

RMSE = 1.276125

Multiple regression results:

R-squared = 0.850777

Adjusted R-squared = 0.848912

F = 456.111097

Model p-value = 1.454898041e-258

RMSE = 1.247020

Strongest standardized predictor = G2 (Beta = 0.797983)

Interpretation:

Simple regression estimates the relationship between G2 and G3.

Multiple regression estimates each predictor effect while holding the other predictors constant.

Review p-values, confidence intervals, standardized beta coefficients, residual diagnostics, VIF, leverage, and Cook's distance before final reporting.

Charts generated: 10

Model comparison

model	outcome	predictors	observatio	number_o	r_squared	adjusted_r	f_statistic	f_p_value	aic	bic	rmse	mae	residual_s	durbin_wa	decision_alp
Simple lin	G3	G2	649.00000	1.00000	0.84373	0.84349	3493.2815	0.00000	2162.2708	2171.2216	1.27612	0.80802	1.27810	1.84640	Model is stat
Multiple li	G3	G1, G2, st	649.00000	8.00000	0.85078	0.84891	456.11110	0.00000	2146.3246	2186.6035	1.24702	0.78002	1.25576	1.86154	Model is stat

Regression coefficients

model	term	coefficient_b	standard_error	t_value	p_value	ci95_lower	ci95_upper	standardized_b	absolute_stand	decision_alpha_0
Simple linear r	Intercept	0.12197	0.20559	0.59324	0.55323	-0.28174	0.52568			Not statistically s
Simple linear r	G2	1.01849	0.01723	59.10399	0.00000	0.98465	1.05233	0.91855	0.91855	Statistically signi
Multiple linear	Intercept	-0.50115	0.77395	-0.64753	0.51752	-2.02094	1.01863			Not statistically s
Multiple linear	G1	0.14340	0.03667	3.91027	0.00010	0.07138	0.21541	0.12185	0.12185	Statistically signi
Multiple linear	G2	0.88481	0.03437	25.74397	0.00000	0.81732	0.95230	0.79798	0.79798	Statistically signi
Multiple linear	studytime	0.09663	0.06209	1.55630	0.12013	-0.02529	0.21856	0.02481	0.02481	Not statistically s
Multiple linear	failures	-0.23536	0.09523	-2.47139	0.01372	-0.42237	-0.04835	-0.04322	0.04322	Statistically signi
Multiple linear	absences	0.02276	0.01092	2.08457	0.03750	0.00132	0.04420	0.03270	0.03270	Statistically signi
Multiple linear	age	0.02269	0.04367	0.51950	0.60359	-0.06306	0.10844	0.00855	0.00855	Not statistically s
Multiple linear	Medu	-0.04495	0.05794	-0.77585	0.43812	-0.15872	0.06882	-0.01579	0.01579	Not statistically s
Multiple linear	Fedu	0.02203	0.05932	0.37133	0.71052	-0.09445	0.13850	0.00750	0.00750	Not statistically s

Regression ANOVA

model	source	sum_of_squares	df	mean_square	f_value	p_value
Simple linear regressio	Regression	5706.37384	1.00000	5706.37384	3493.28157	0.00000
Simple linear regressio	Residual	1056.89272	647.00000	1.63353		
Simple linear regressio	Total	6763.26656	648.00000			
Multiple linear regressi	Regression	5754.03298	8.00000	719.25412	456.11110	0.00000
Multiple linear regressi	Residual	1009.23359	640.00000	1.57693		
Multiple linear regressi	Total	6763.26656	648.00000			

Multiple-regression VIF

predictor	vif	tolerance	diagnostic
G1	4.16483	0.24011	Acceptable
G2	4.12080	0.24267	Acceptable
studytime	1.09008	0.91736	Acceptable
failures	1.31161	0.76242	Acceptable
absences	1.05519	0.94770	Acceptable
age	1.16273	0.86004	Acceptable
Medu	1.77555	0.56321	Acceptable
Fedu	1.74917	0.57170	Acceptable

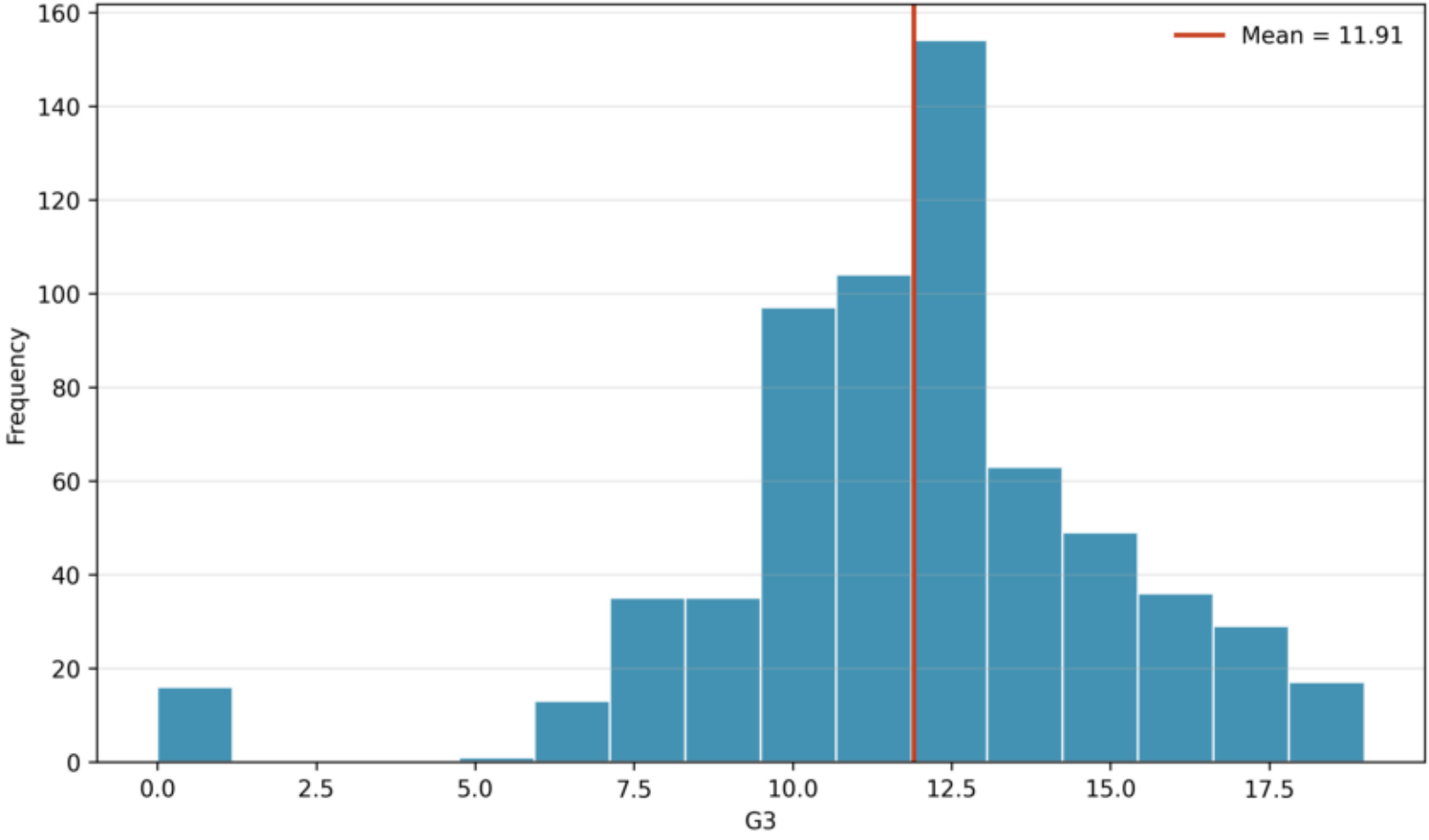
Diagnostic tests

model	diagnostic	statistic	p_value	interpretation
Simple linear regression	Shapiro-Wilk residual normality	0.71085	0.00000	Potential non-normality
Simple linear regression	Jarque-Bera residual normality	10482.94671	0.00000	Potential non-normality
Simple linear regression	Breusch-Pagan heteroscedasticit	15.75877	0.00007	Potential heteroscedasticity
Simple linear regression	Breusch-Pagan F test	16.10116	0.00007	Potential heteroscedasticity
Simple linear regression	Durbin-Watson residual indepen	1.84640		Values near 2 indicate approximat
Multiple linear regression	Shapiro-Wilk residual normality	0.75679	0.00000	Potential non-normality
Multiple linear regression	Jarque-Bera residual normality	9992.99990	0.00000	Potential non-normality
Multiple linear regression	Breusch-Pagan heteroscedasticit	35.72281	0.00002	Potential heteroscedasticity
Multiple linear regression	Breusch-Pagan F test	4.65992	0.00001	Potential heteroscedasticity
Multiple linear regression	Durbin-Watson residual indepen	1.86154		Values near 2 indicate approximat

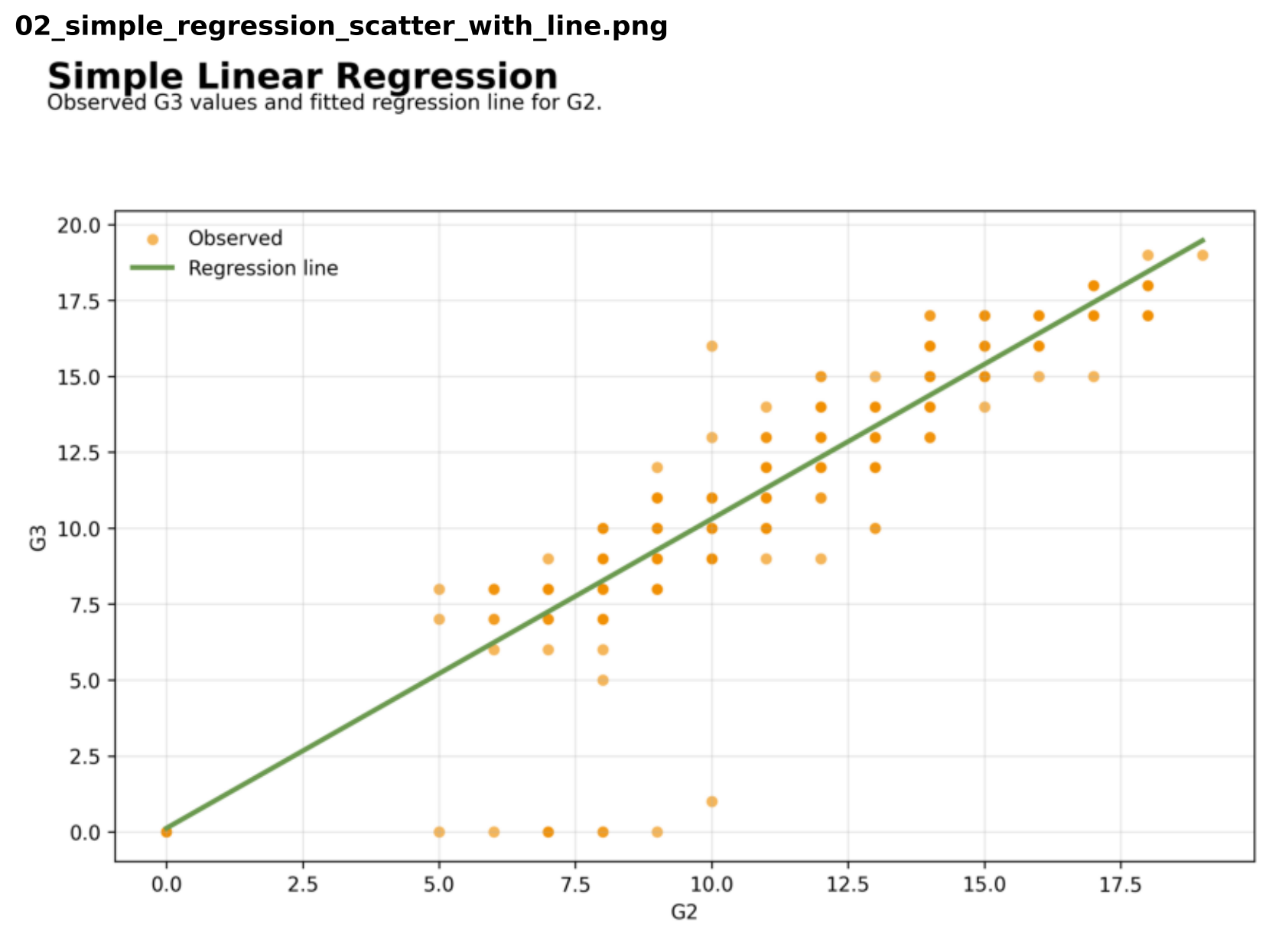
01_regression_in_python_outcome_distribution.png

Outcome Distribution

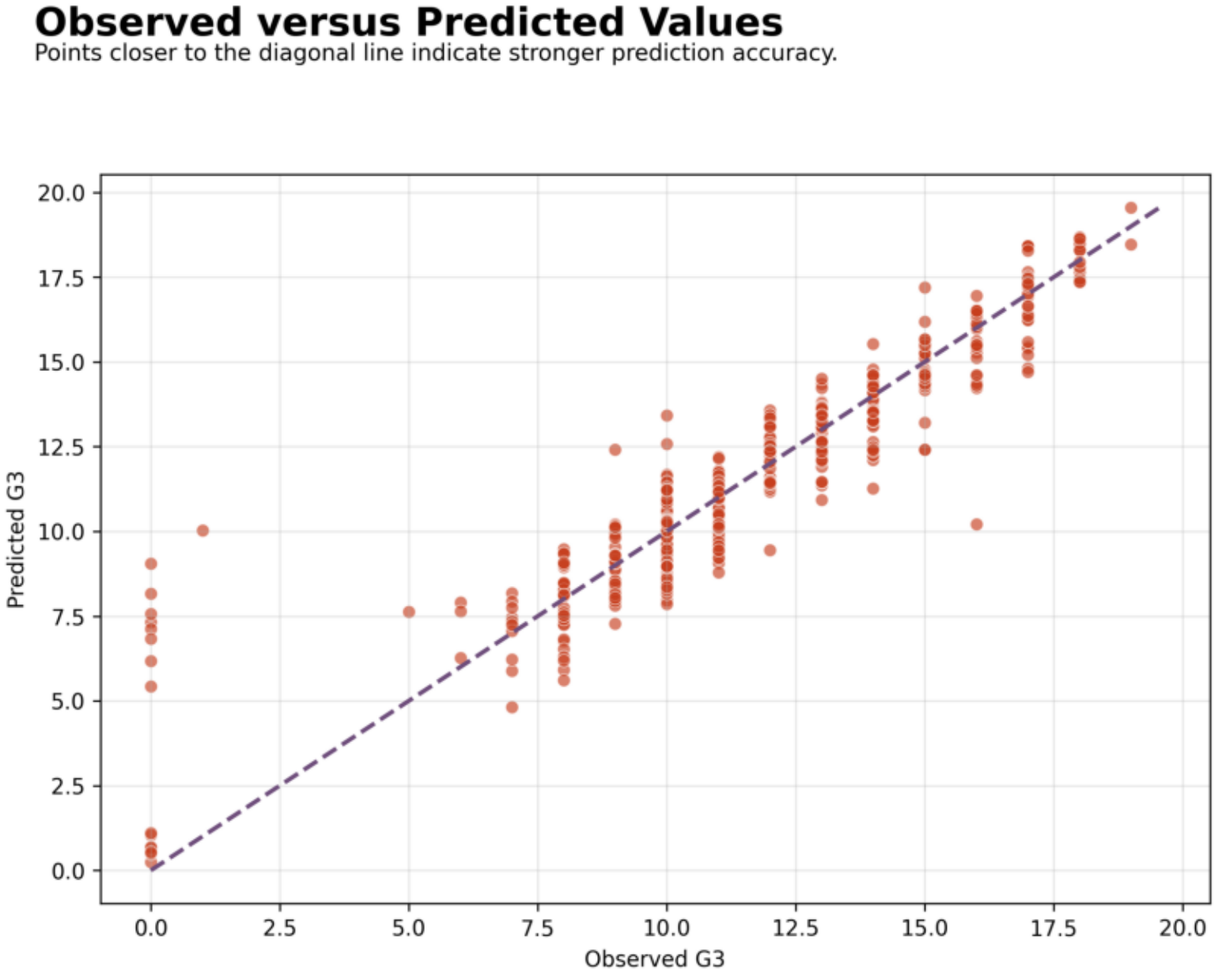
Distribution of the dependent variable used in the regression analyses.



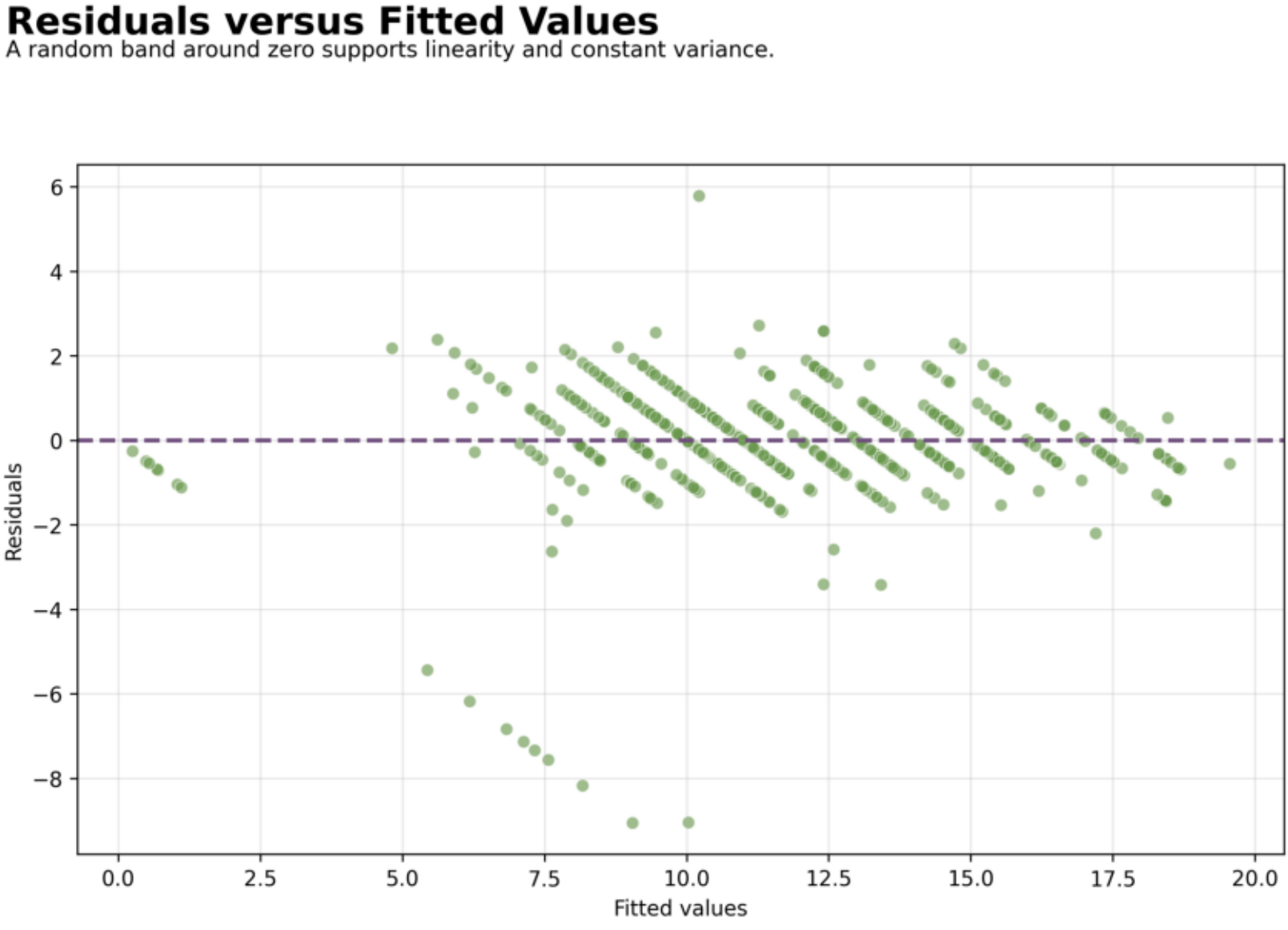
02_simple_regression_scatter_with_line.png



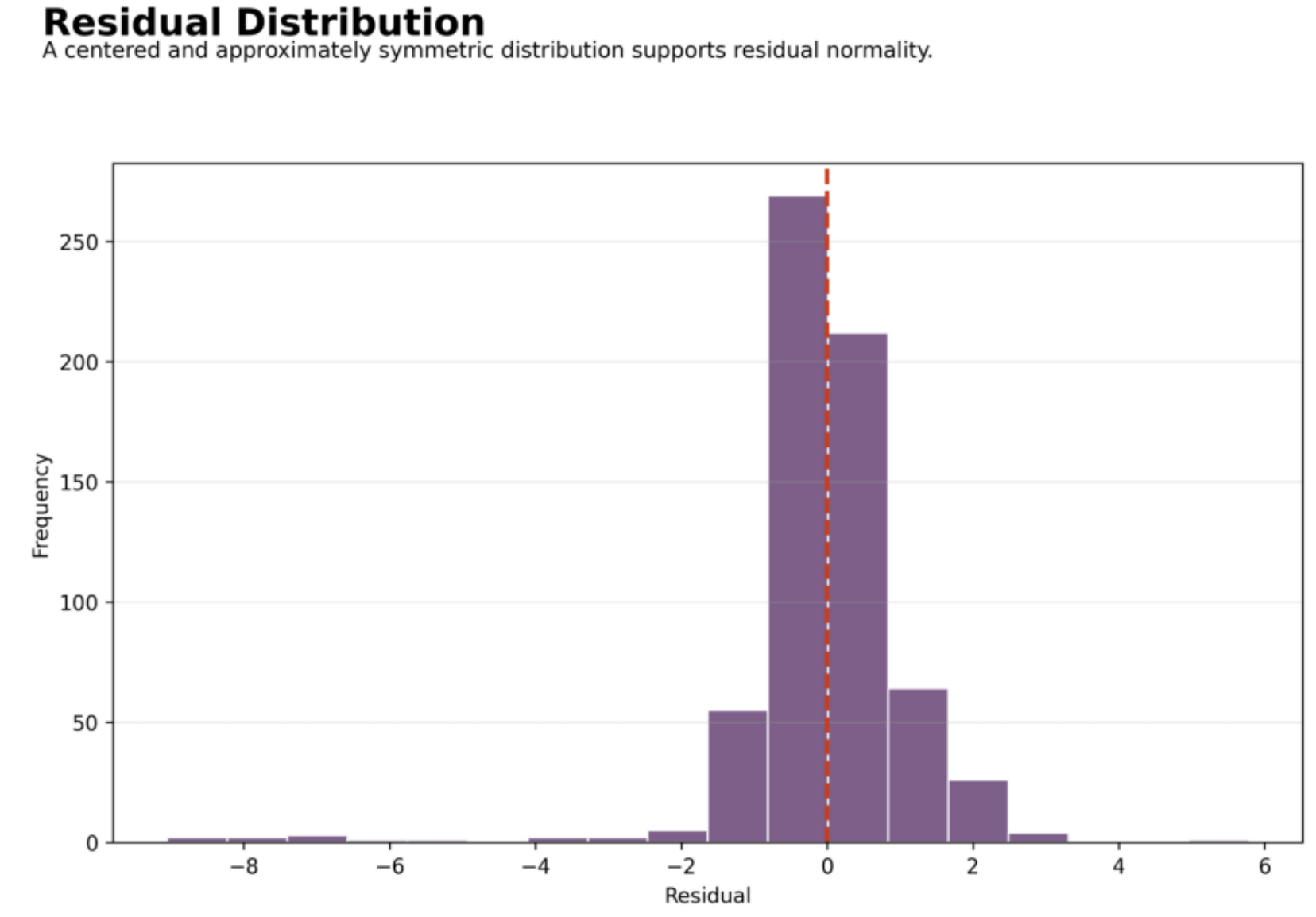
03_multiple_regression_observed_vs_predicted.png



04_multiple_regression_residuals_vs_fitted.png



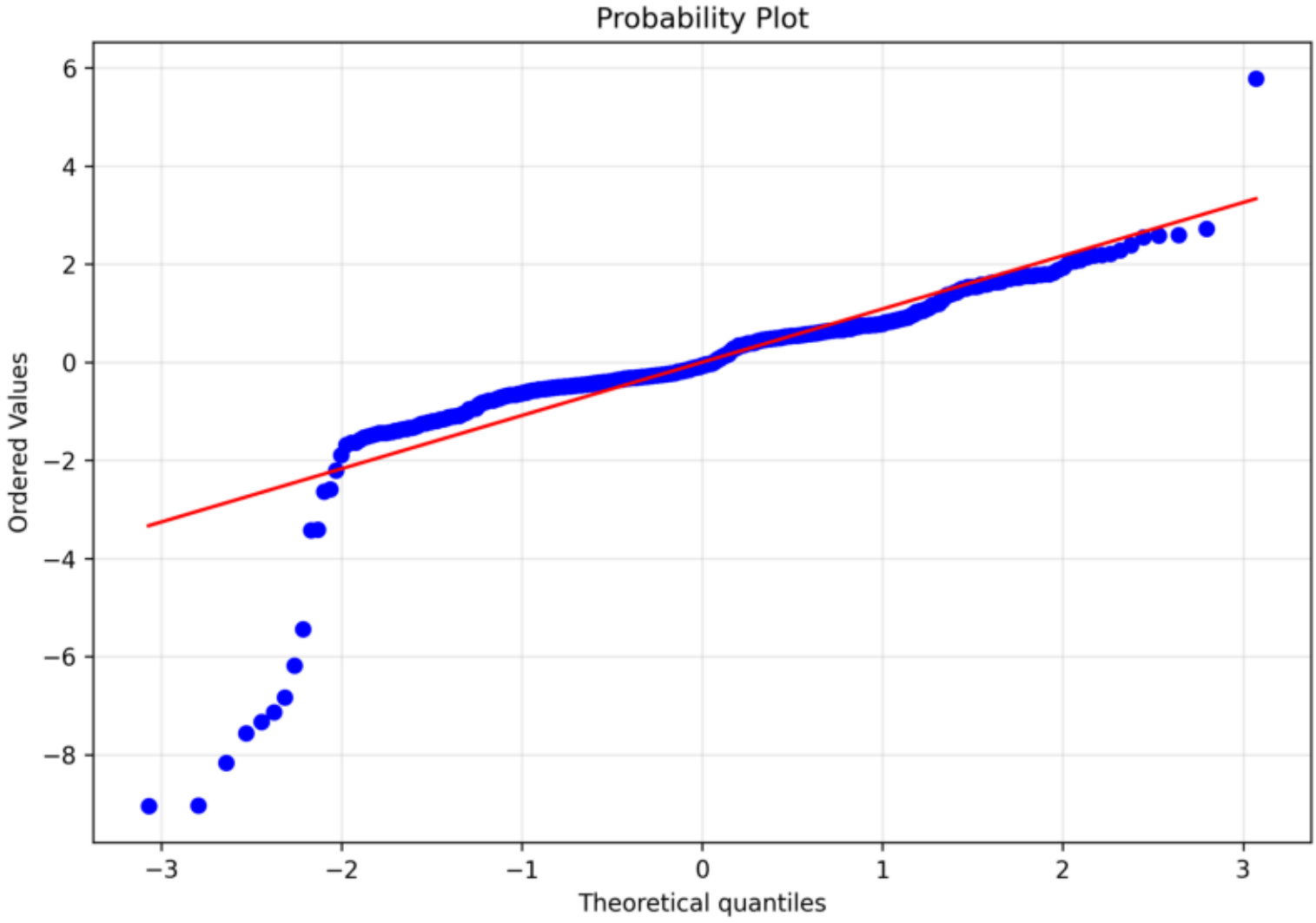
05_multiple_regression_residual_histogram.png



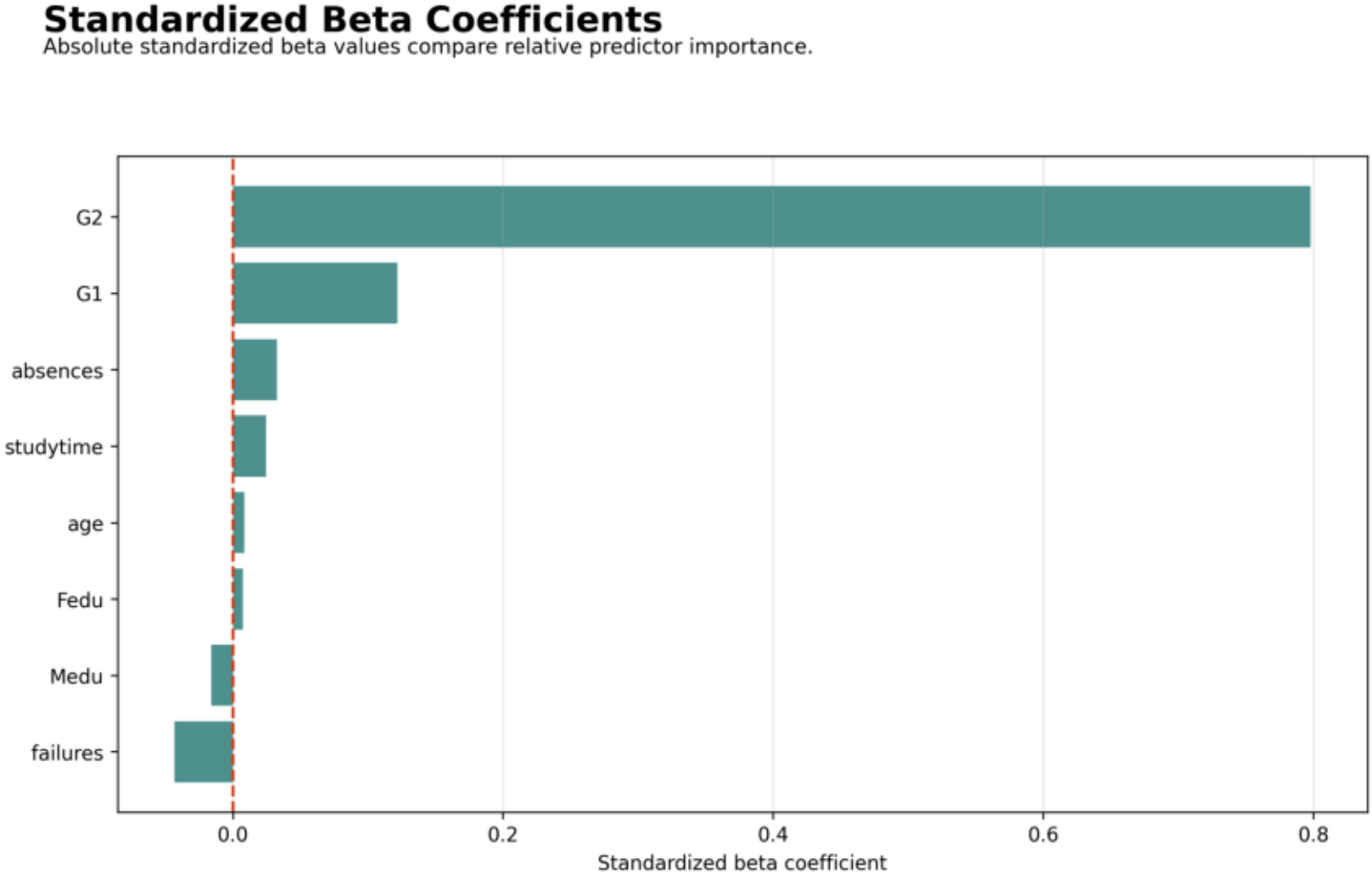
06_multiple_regression_residual_qq_plot.png

Q-Q Plot of Residuals

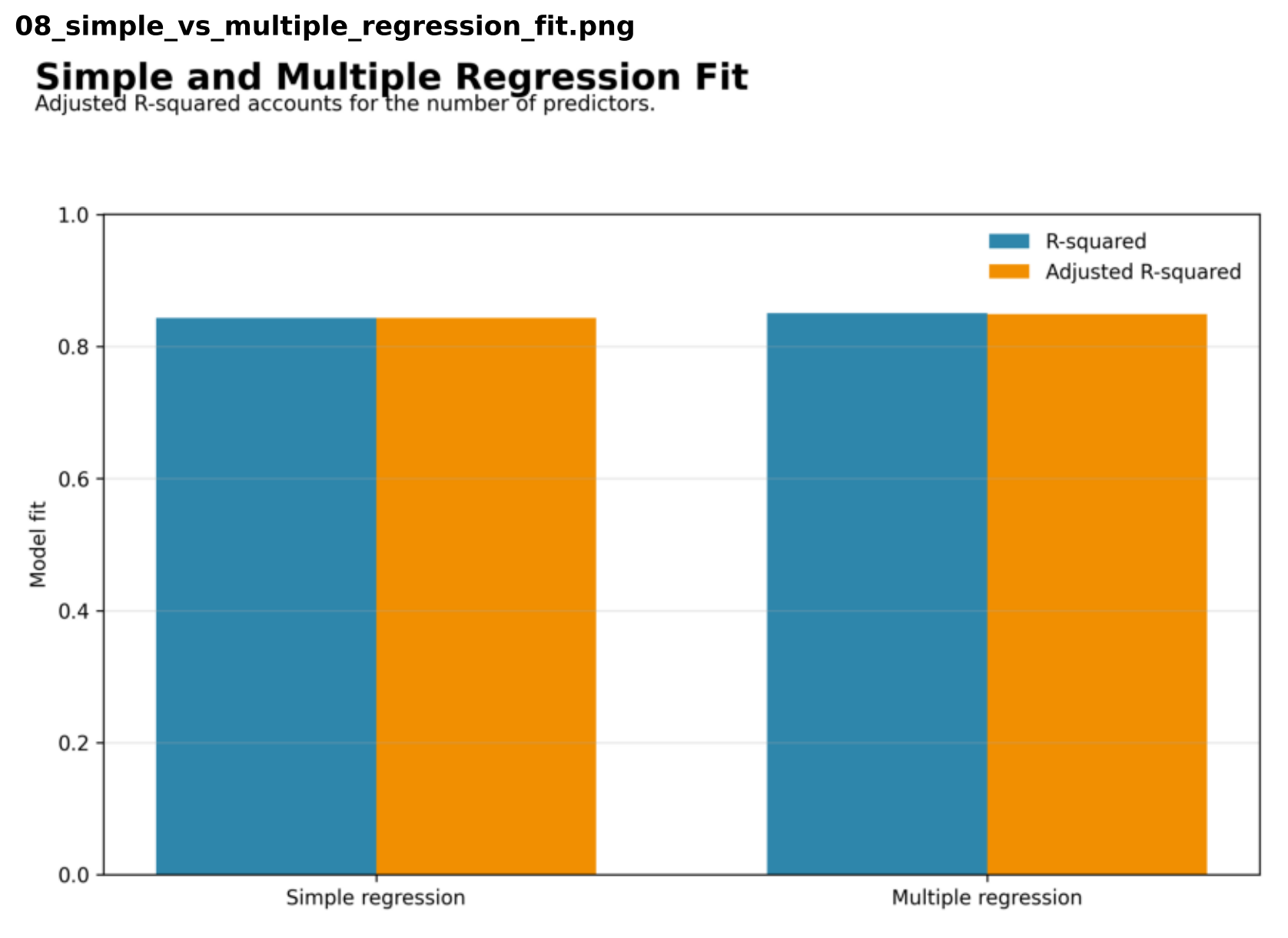
Points close to the reference line support approximate normality.



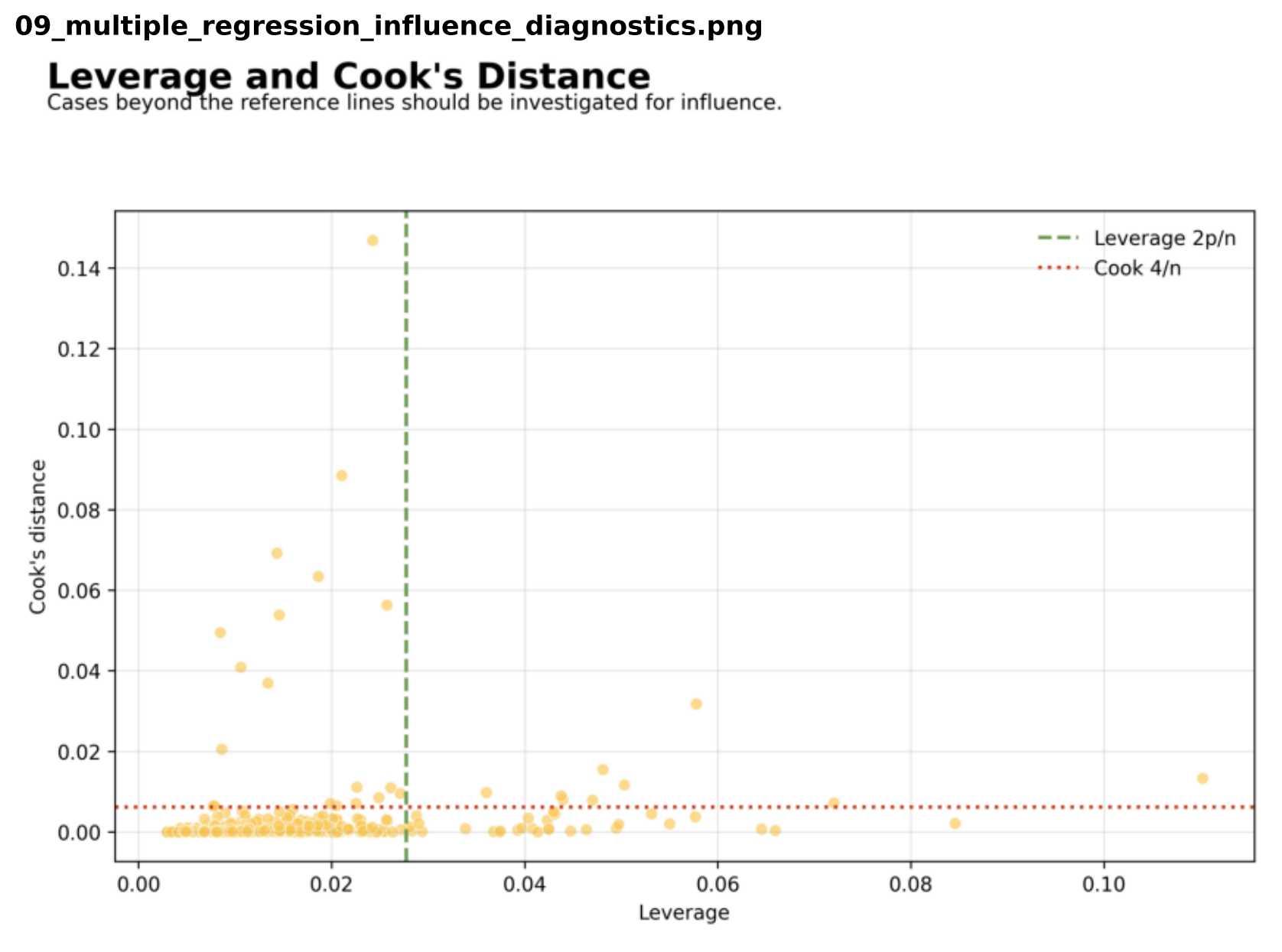
07_multiple_regression_standardized_betas.png



08_simple_vs_multiple_regression_fit.png



09_multiple_regression_influence_diagnostics.png



10_regression_variable_correlation_heatmap.png

Regression Variable Correlation Matrix

Strong predictor correlations may affect coefficient stability.